A Supplementary Material

A.1 Dataset Details

Job advertisement dataset. First, we use a job advertisement collected from Wantedly[2]. This dataset includes 52,914 job advertisement articles written in Japanese from June 2012 to November 2018. Each article contains manually created attributes including a headline, key phrase and category. This dataset is characterized by the headlines, key phrases and categories containing common features, such as occupations. For example, in Figure 1, the word "Engineer" is commonly noted in the headline, key phrase and category. We split the dataset into train, test and valid sets as presented in Table 8. Articles written by the same company belong to the same set to avoid leakage.

The key phrases of the articles have abstractness. For example, the model is expected to generate the key phrase "data scientist" based on the description "engineers who have an essential ability to conduct data analysis." Hence, to generate these key phrases, the abstractive summarization method is more suitable than the extractive summarization method.

The CNN-DM dataset. We also evaluate on the CNN-DailyMail (CNN-DM) dataset, which was originally created for machine reading comprehension tasks (Hermann et al., 2015). The original CNN-DM dataset and the raw HTMLs of the articles are distributed on their websites [3]. Extractive summarization (Cheng and Lapata, 2016) and abstractive summarization studies (See et al., 2017) use this CNN-DM dataset, and they distributed the script [4] to preprocess the dataset.

To evaluate multi-sentence summarization, headline generation and article classification tasks, we build the modified CNN and DailyMail datasets by extracting the headlines and categories from the raw HTMLs of the articles. This extraction procedure was executed as follows:

Dataset splitting. We extracted the body texts and multi-sentence summaries from the original CNN and DailyMail articles. We used the scripts of See et al. (2017) to extract and split the articles from the Raw HTMLs. We tokenized the sentences by employing the Stanford CoreNLP parser (Manning et al., 2014). Note that 114 articles were omitted as these articles contained no body texts.

	Job Ads	CNN	DailyMail
Train data	48,406	90,018	196,958
Test data	3,748	1,090	10,385
Valid data	760	1,220	12,148
Number of categories	18	44	21

Table 8: Statistics of the job advertisement dataset (Job Ads), CNN and DailyMail datasets.

	CNN	DailyMail
Number of articles	92,328	219,491
Average headline length	8.8	17.9
Average sentences per summary	3.6	3.9
Average summary length	56.1	66.2
Average article length	763.1	801.8

Table 9: Statistics of the CNN and DailyMail datasets.

Extraction of headlines. We extracted the sentences in <title> tags, and treated them as headlines. Almost all the headlines in the dataset were composed of suffixes, such as " - CNN.com" in common; therefore, they were removed from the headlines. Table 9 presents the statistics of the CNN and DailyMail datasets. The average length of the headlines in the CNN dataset is 8.8 words, whereas that in the DailyMail dataset is 17.9 words. These characteristics result in the difference between the ROUGE scores of the headline generation in both datasets (Table 4).

Extraction of Categories. We also extracted categories of the articles from the raw HTMLs. For the CNN dataset, we extracted the categories from the <meta> tags. Meanwhile, for the DailyMail dataset, we extracted categories from the URL of the articles. Note that 152 articles were omitted as we could not extract the categories.

Table 9 and Table 10 show the statistics of the CNN and DailyMail datasets. From these statistics, it is noticeable that the CNN and DailyMail datasets have different taxonomies, and that the DailyMail dataset has more imbalanced categories. As a result, the classification accuracy scores for the DailyMail dataset is significantly higher than that of the CNN dataset (Table 4).

Figure 6 and 7 are examples of articles of the CNN and DailyMail datasets, respectively.

[2] Wantedly is one of the most popular job-matching web services in Japan. https://www.wantedly.com

[3] https://cs.nyu.edu/~kcho/DMQA/

[4] https://github.com/abisee/cnn-dailymail

	CNN		DailyMail	
	Category	Freq.	Category	Freq.
1st	World	24.1 %	News	67.0 %
2nd	US	13.4 %	Sport	16.2 %
3rd	Politics	7.7 %	Female	6.0 %
4th	Opinion	7.6 %	Science&Tech	5.1 %
5th	Crime	6.1 %	Health	3.0 %

Table 10: Top five categories shown of the CNN and DailyMail datasets.

A.2 Implementation Details

Job advertisement dataset. Our model contains 300-dimensional word embeddings and 256-dimensional hidden layers. We use mecab-ipadic-NEologd (Sato et al., 2017) for word segmentation, and FastText (Bojanowski et al., 2017) as the pretrained word embeddings. The vocabulary size is set to 12,000, and the word embedding layer is shared among the encoder and task-specific decoders. We train the model using Adam (Kingma and Ba, 2015) optimizer at a learning rate of 0.001. We truncate the input articles to 300 words, while the output lengths of the headlines and key phrases are limited to 32 and 8, respectively.

We train the model for 8 epochs. We set the hyperparameters of the loss function as follows: $\lambda^1_{const} = 5.0, p^1_{th} = 0.0$ for the classification, $\lambda^2_{const} = 1.0, p^2_{th} = 1.0$ for the key phrase generation and $\lambda^3_{const} = 1.0, p^3_{th} = 2.0$ for the headline generation task. This implies that we first train the model for the classification task. Then, the trained information is transferred to more difficult tasks, including headline and key phrase generation tasks. In addition, we set $\lambda^2_{cov} = 0.2, \lambda^3_{cov} = 0.2$ and $\lambda^{all}_{hcl} = 0.1$.

The CNN-DM dataset. Following See et al. (2017), we set the size of the word embedding layer as 128 dimensions, and the hidden layers as 256 dimensions. We use a vocabulary of 50,000 words for both the encoder and task-specific decoders. We adopt AdaGrad (Duchi et al., 2011) optimizer with a learning rate 0.15 and an initial accumulator value of 0.1.

We train the model for 47 epochs for the CNN dataset, and 25 epochs for the DailyMail dataset. Following See et al. (2017), we apply early stopping for training the headline generator and classifier on the baseline, avoiding overfitting. We set the hyperparameters of the loss function as follows: $\lambda^1_{const} = 0.1, p^1_{th} = 0.0$ for the classification, $\lambda^2_{const} = 1.0, p^2_{th} = 1.0$ for the headline generation and $\lambda^3_{const} = 1.0, p^3_{th} = 2.0$ for the multi-sentence summarization task. In addition, we set $\lambda^2_{cov} = 0.2, \lambda^3_{cov} = 1.0$ and $\lambda^{all}_{hcl} = 1.0$.

We apply an additional scheduling strategy to avoid destabilization of the training process. Following See et al. (2017), we activate the coverage loss during the last epochs. We avoid simultaneously adding both coverage loss and hierarchical consistency loss to the overall loss function. We inactivate the hierarchical consistency loss while the coverage loss is activated. In addition, we separately train each task-specific decoder while the coverage loss is activated, because the simultaneous combination of multiple coverage losses destabilizes the training process.

We train the model on a GTX-1080Ti GPU. The training phase lasts for approximately 30 hours for the job advertisement dataset, and approximately 5.5 days for the CNN and DailyMail datasets. We set the mini-batch size to 16, dropout rate to 0.5 and the beam search size to 5. We use a gradient clipping method with a maximum gradient norm of 2 per tasks. We use an approximate randomization test (Noreen, 1989) to evaluate the statistically significance (sample size is 10,000).

Headline: the mentally ill – jailed and desperate for help

Multi-Sentence Summary:
<l> mentally ill inmates in miami are housed on the “ forgotten floor ” </l>
<l> judge steven leifman says most are there as a result of “ avoidable felonies ” </l>
<l> while cnn tours facility , patient shouts : “ i am the son of the president ” </l>
<l> leifman says the system is unjust and he ’s fighting for change . </l>

Category: News

Description Text (Truncated):
editor ’s note : in our behind the scenes series , cnn correspondents share their experiences in covering news and analyze
the stories behind the events . here , soledad o’brien takes users inside a jail where many of the inmates are mentally
ill . an inmate housed on the “ forgotten floor , ” where many mentally ill inmates are housed in miami before trial .
miami , florida -lrb- cnn -rrb- – the ninth floor of the miami-dade pretrial detention facility is dubbed the “ forgotten floor .”
here , inmates with the most severe mental illnesses are incarcerated until they ’re ready to appear in court .
most often , they face drug charges or charges of assaulting an officer – charges that judge steven leifman says are usually
“ avoidable felonies . ” he says the arrests often result from confrontations with police . mentally ill people often wo n’t
do what they ’re told when police arrive on the scene – confrontation seems to exacerbate their illness and they become
more paranoid , delusional , and less likely to follow directions , according to leifman . so , they end up on the ninth floor
severely mentally disturbed , but not getting any real help because they ’re in jail . we toured the jail with leifman .
he is well known in miami as an advocate for justice and the mentally ill . even though we were not exactly (...)

Figure 6: Example of article of the CNN dataset with a headline and category.

Headline: barefoot bandit colton harris-moore pleads guilty to theft of airplanes , boats and cars

Multi-Sentence Summary:
<l> colton harris-moore , 20 , changed plea to guilty . </l>
<l> he lived on the run from april 2008 to july 2010 . </l>
<l> allegedly committed 100 + crimes , some while barefoot . </l>

Category: News

Description Text (Truncated):
by . daily mail reporter . last updated at 11:05 pm on 17th june 2011 .
pleaded : colton harris-moore , aka the barefoot bandit , pleaded guilty to a string of sensational thefts .
the young washington state man who gained international notoriety during a two-year run from the law in stolen boats ,
cars and planes has pleaded guilty to seven charges in the ‘ barefoot bandit ’ case . colton harris-moore , 20 ,
entered the plea in federal court friday morning , reversing a not guilty plea made last week and ending
the latest chapter in his fugitive saga . he could receive between 5 1/4 and 6 1/2 years in prison when he ’s sentenced
in october , said harris-moore ’s attorney , john henry browne . federal prosecutors were to give details of
the plea deal reached with the 20-year-old ’s lawyers after the friday morning hearing .
the two sides had been negotiating whether harris-moore could participate in book or movie deals ,
with proceeds used to repay victims . his lawyers have said restitution could total about $ 1.3 million .
prosecutors have said harris-moore hopscotched his way across the united states , frequently crash-landing planes
in rural areas and stealing cars from parking lots at small airports . (...)

Figure 7: Example of article of the DailyMail dataset with a headline and category.